DOCUMENT RESUME

ED 410 241 TM 026 978

AUTHOR Shen, Linjun
TITLE Quantifying Item Dependency by Fisher's Z.
PUB DATE Mar 97
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Estimation (Mathematics); Item Response Theory; *Reliability; Statistical Analysis; *Test Items; *Validity
IDENTIFIERS *Fishers Z Transformation; *Item Dependence

ABSTRACT

Three aspects of the usual approach to assessing local item dependency, Yen's "Q" (H. Huynh, H. Michaels, and S. Ferrara, 1995), deserve further investigation. Pearson correlation coefficients do not distribute normally when the coefficients are large, and thus cannot quantify the dependency well. In the second place, the accuracy of item response theory person ability estimates is always relative to the standard error of the estimation, and in the third place, Yen's approach does not have a criterion to determine the significance level of the index of dependency between a pair of items (Q3). This paper addresses these concerns by proposing Fisher's "Z" as the index for item dependency. It compares the statistical properties of Fisher's Z with Yen's Q and weights the observed residuals by the standard error of person ability. Items for the study were from the second administration of part of a national medical licensing examination in 1995 taken by 361 first-time test takers. Fisher's Z accurately estimated the zero mean dependency for stand-alone items and successfully differentiated stand-alone items from clustered items based on the degree of dependency. This suggests that Fisher's Z is a valid, reliable, and sensitive index for quantifying item dependency. Two appendixes provide a sample of stand-alone items and a sample of item cluster. (Contains four figures and three references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 410 241

Quantifying Item Dependency by Fisher's Z

Linjun Shen

National Board of Osteopathic Medical Examiners

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Linjun Shen

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as
received from the person or organization
originating it.
Minor changes have been made to
improve reproduction quality.
Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

**Paper presented to the Annual Meeting of the
American Educational Research Association
Chicago, March 1997**

TM026978

BEST COPY AVAILABLE

Quantifying Item Dependency by Fisher's Z

One of the commonly used methods to assess local item dependency is Yen's Q_3 (Huynh, Michaels, & Ferrara, 1995) . This approach computes residuals between examinee's observed score and IRT- derived person ability score on each item for all persons. Q_3, the index of dependency between a pair of items, is the Pearson correlation coefficient between the residual scores on that pair of items. For item clusters, the mean of all possible inter-item Q_3 within a cluster is considered as the index of the dependency for that cluster (Yen, 1984, 1993). Three aspects of this approach need further investigation. First, Pearson correlation coefficients do not distribute normally especially when the coefficients are large. Therefore, Pearson coefficient cannot quantify the dependency well. Second, the accuracy of IRT person ability estimates is always relative to the standard error of the estimation. The Q_3, relying on raw residual scores between observed score and person ability, failed to take into account the accuracy of person measure. Third, Yen's approach does not have a criterion to determine the significance level of Q_3 for clustered items. Determination of item dependency remains arbitrary.

The purpose of this paper was to address the above three concerns by proposing Fisher's Z as the index for item dependency and to compare the statistical properties of Fisher's Z with Yen's Q_3. Fisher's Z normalized Pearson correlation coefficients, therefore provided a better-scaled index. To acknowledge the effect of error, this study weighted the observed residuals by the standard error of person ability. This paper established a criterion to determine the significance of Fisher's Z by using the level of dependency among stand-alone items as the reference. A Fisher's Z for a pair of clustered items would be considered significant if it was at least two standard deviations above the mean of Fisher's Z among stand-alone items in the same exam. The rationale was: the distribution of Fisher's Z among all stand-alone items must have a mean of zero. If a Fisher's Z was larger than two standard deviations of the Fisher's Z for stand-alone items, either it was due to the random effect or the two items were not independent stand-alone items.

Methods

Subjects and instruments

Items in this study were selected from the second administration of the third part of a national medical licensing examinations in 1995. This four-book exam had 770 multiple-choice items. The KR-20 reliability coefficient was .94. 414 items were stand-alone one-best-answer items as illustrated in Appendix 1, the other 356 items were clustered into 130 item sets with number of items per set varying between 2-8. As Appendix 2 shows, items in a cluster shared a common clinical presentation, therefore, they tended to be dependent with each other.

416 medical students with at least six months of postgraduate medical education took the exam. For reliable results, this study only used 361 first-time takers.

Item selection

All 356 items in clusters were analyzed. Within-cluster Fisher's Z matrix were computed for each of 130 clusters. The total number of unique Fisher's Z for clustered items was 378.

To obtain a criterion distribution of Fisher's Z, five groups of stand-alone one-best-answer items were randomly selected from the exam. Each group had 12 items and consequently yielded a correlation matrix with 66 unique Fisher's Z. Five 12 x 12 matrices gave 330 Fisher's Z which constructed the Fisher's Z distribution for stand-alone items.

Computation of Fisher's Z

Rasch Model was the foundation of this paper. All items and persons were analyzed by this Model. The computation of Fisher's Z took three steps:

1) Obtain the Rasch Model-based standardized residuals for each observation of person n on item i :

$$d_{ni} = \frac{Observed_{ni} - Expected_{n}}{SE_{n}}$$

2) Correlate the standardized residuals d_{ij} for all pairs of items i,j across all N persons

3) Compute Fisher's Z to normalize Pearson correlation r_{ij} from the step 2:

$$Z_{ij} = \frac{1}{2}\log\left(\frac{1+r_{ij}}{1-r_{ij}}\right)$$

with error of $r_{ij}/2[N-1]$.

Results

Distributions of Fisher's Z

As the histogram in Figure 1 shows, the distribution of 330 unique Fisher's Z for stand-alone items was normal. The mean of the distribution was -.004 with a standard deviation of .06. This distribution suggested that the probability of having a Fisher's Z of .12 (two standard deviations above the mean) between any of two stand-alone items was only .05 or less. So, .12 was considered as the critical value of significant Fisher's Z for clustered items.

In contrast, the distribution of Fisher's Z for clustered items in Figure 2 was skewed with a

mean of .20 and a standard deviation of 0.32. 164 (43%) of Fisher's Z for clustered items were significant. 68 out of 130 clusters (52%) had a mean Fisher's Z greater than .12. On the other hand, 45 cases (35%) did not have any pair of items significantly dependent.

Comparison between Fisher's Z and Q_3

To evaluate the effect of standardization of residuals by the standard error of person ability, Fisher's Z based on observed residuals was plotted against corresponding Fisher's Z based on standardized residuals. The plot almost fell into an identity line and the correlation was .98.

The effects of the normalization by Fisher's Z for the Q_3 was demonstrated by a scatter plot of Fisher's Z against the corresponding Q_3. As Figure 3 illustrates, the plot did not fall into an identity line and the relationship was not linear. The Figure 3 indicated that when the Q_3 was relatively small, the plot followed a straight line. However, if the Q_3 was above .60, an equal increment in the Q_3 would correspond to an increasingly larger increment in Fisher's Z. This relation pattern indicates that the Q_3 has the tendency to reduce the degree of dependency when the Q_3 is large. Transformation from Q_3 into Fisher's Z avoided this problem.

Discussion

Fisher's Z accurately estimated the zero mean dependency for stand-alone items and successfully differentiated stand-alone items from clustered items based on the degree of dependency. This suggests that Fisher's Z is a valid, reliable, and sensitive index for quantifying item dependency.

Compared with Yen's Q_3, Fisher's Z has better statistical properties due to its standardization of residuals, normalization of the Q_3, and a practical criterion of significance. Although the comparison of the standardized residuals with the observed residuals in this study did not show practically meaningful differences, the necessity and advantage of standardization by measurement error is still conceptually obvious. The exam analyzed in this study was a well-established standardized written test, contained a large amount of well-written multiple-choice items, therefore, the person ability was estimated with a good and consistent precision. This explains why the discrepancy between the observed and standardized residuals was trivial. However, if the items are open to examinees' free expression, subject to relatively subjective scoring, and the number of items in an exam is smaller, as in many performance tests, person ability would be estimated with a large standard error. Moreover, the variation of the errors among examinees would be larger. In such situations, standardization of the residuals would have noticeable impact on the size of Fisher's Z. Simulation studies are needed to further investigate this aspect.

Practical significance level of Fisher's Z is a helpful reference in diagnosing item clusters.

The criterion of significant Fisher's Z established by this study was reasonably severe. According to this standard, 57% pairs of clustered items were identified as not significantly dependent. Out of 130 clusters, 45 clusters did not have any pair of items dependent with each other. If .12 is the absolute standard of significant dependency for any written item clusters in medical licensing exams or if the standard of significant Fisher's Z is exam-specific needs further investigation.

The method developed in this study was based on a paper-pencil exam, however, it is applicable to performance tests or computer adaptive tests also, where problems caused by clustered items are more challenging. The advantages of this approach will be more noticeable where exams are short, items do not have a standard format and require free response, and subjective judging are involved.

Reference

Huynh, H., Michaels, H., and Ferrara, S. (1995). Statistical procedures to identify clusters of items with local dependency. A paper presented at the annual meeting of National Council on Measurement in Education, San Francisco, 1995.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

Appendix 1

A sample of stand-alone item

Appendix 2

A sample of item cluster

Figure 1

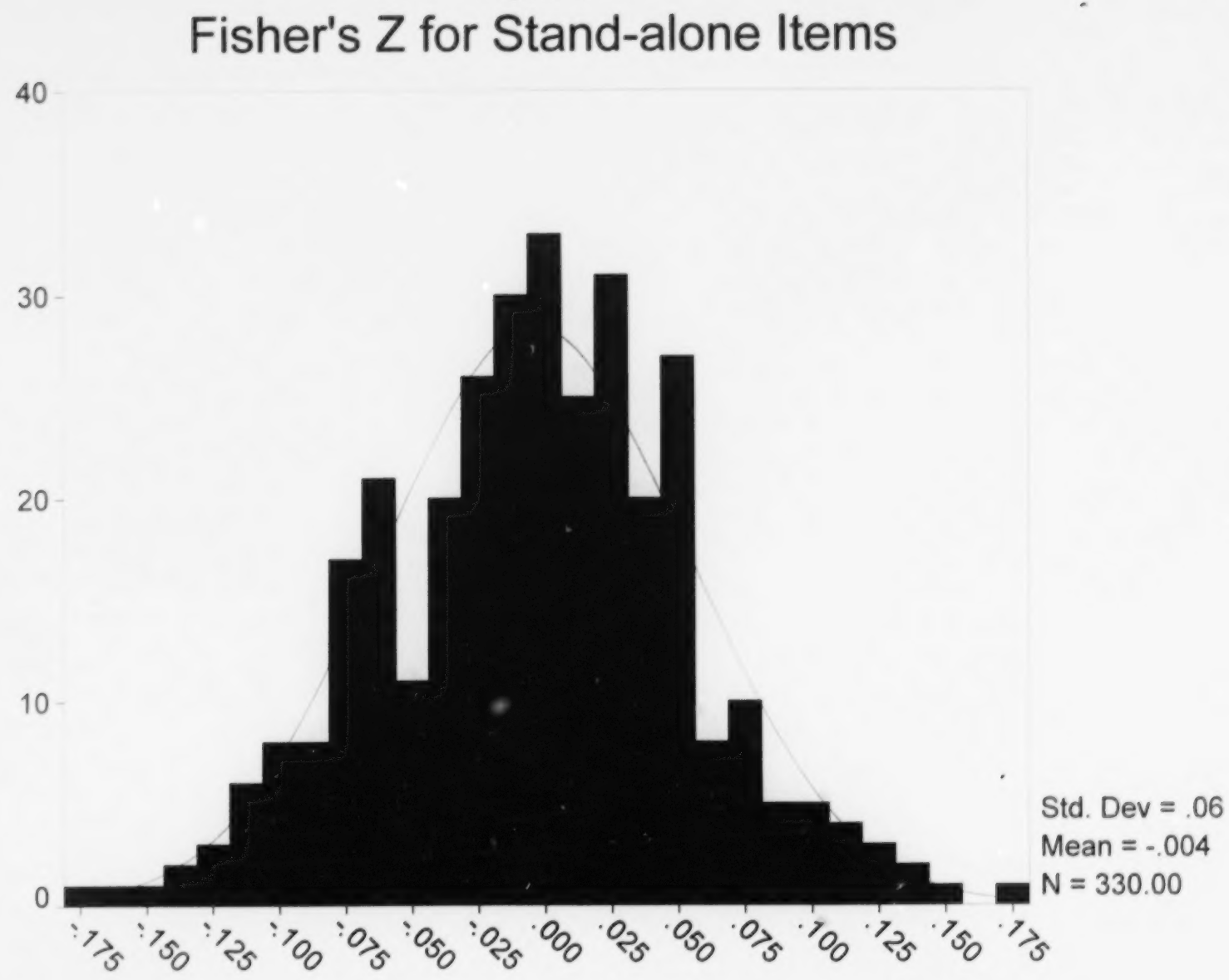
Fisher's Z for Stand-alone Items
40
30
20
10
0
-.175
-.150
-.125
-.100
-.075
-.050
-.025
.000
.025
.050
.075
.100
.125
.150
.175
Std. Dev = .06
Mean = -.004
N = 330.00
Fisher's Z

Figure 2

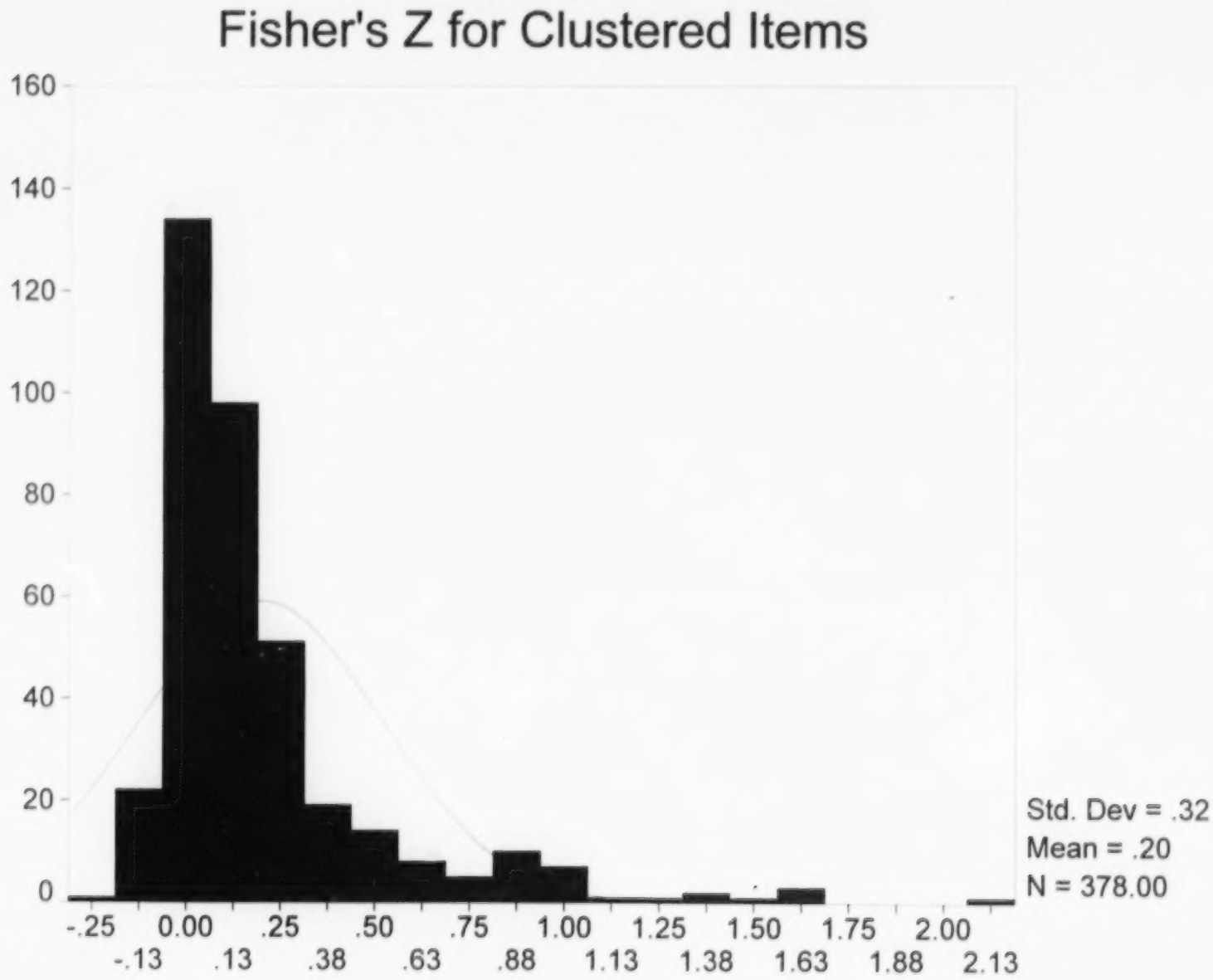
Fisher's Z for Clustered Items
160
140
120
100
80
60
40
20
0
-.25
0.00
.25
.50
.75
1.00
1.25
1.50
1.75
2.00
-.13
.13
.38
.63
.88
1.13
1.38
1.63
1.88
2.13
Std. Dev = .32
Mean = .20
N = 378.00
Fisher's Z

Figure 3

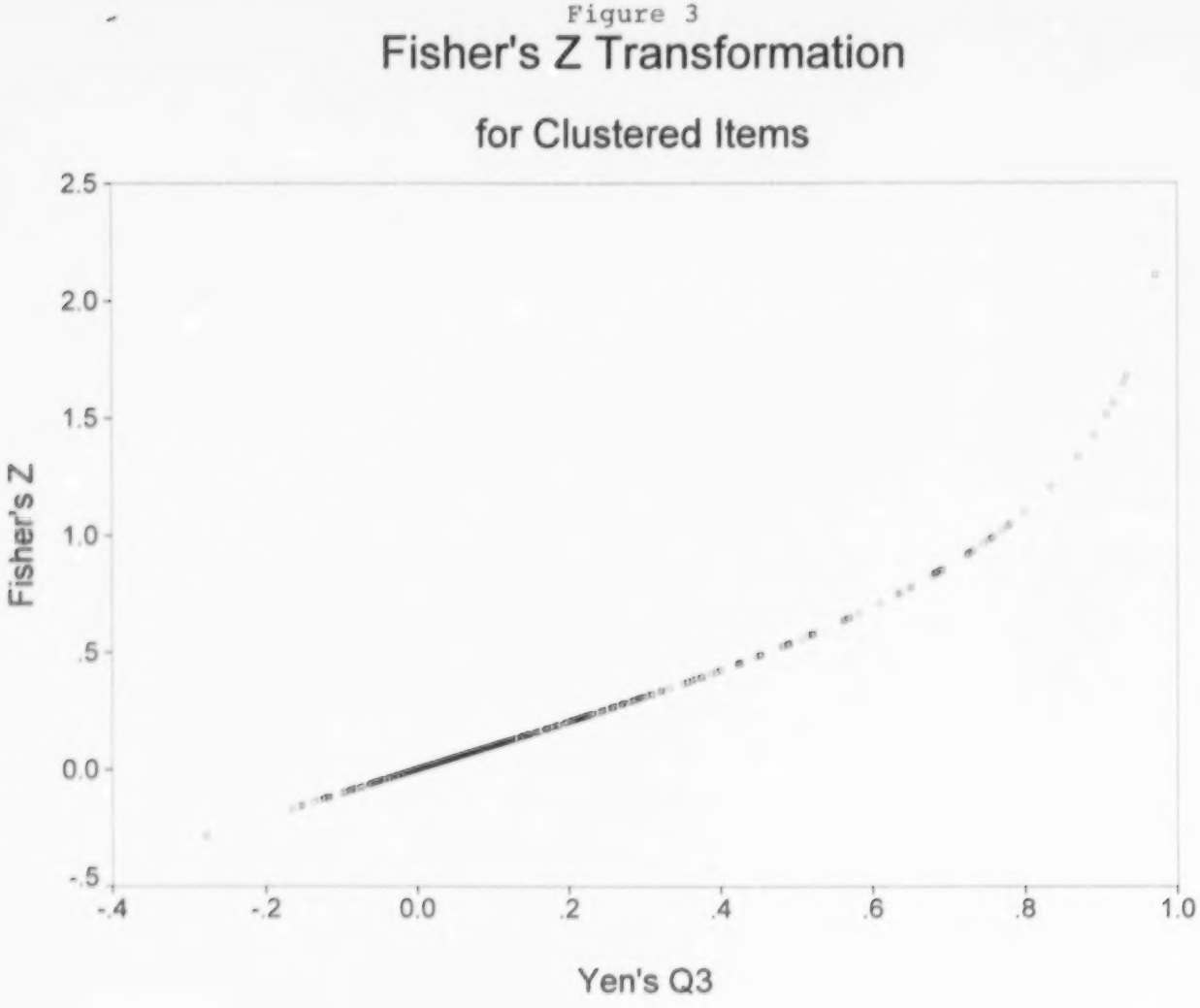
Fisher's Z Transformation
for Clustered Items
Fisher's Z
2.5
2.0
1.5
1.0
.5
0.0
-.5
-.4
-.2
0.0
.2
.4
.6
.8
1.0
Yen's Q3

Figure 3

Fisher's Z Transformation

for Clustered Items